Mise en oeuvre d'un accès ouvert à des ressources linguistiques - bilan du projet Silfide et perspectives.

Laurent Romary, Jean-Marie Pierrel

Mise en oeuvre d'un accès ouvert à des ressources linguistiques — bilan du projet Silfide et perspectives.

Laurent Romary et Jean-Marie Pierrel

Laboratoire Loria

1. Accès normalisé aux ressources linguistiques

La mise en œuvre d'un accès télématique à des ressources linguistiques pose un certain nombre de difficultés éditoriales et techniques qu'il est nécessaire d'appréhender globalement pour éviter les réponses *ad hoc* qui oblitéreraient la pérennité des efforts investis. On peut ainsi mettre en avant quelques aspects essentiels susceptibles de guider la réalisation d'une telle entreprise et envisager dès maintenant un certain nombre de réponses :

- le système doit pouvoir s'appuyer sur des données minimalement normalisées, afin que les efforts de codage et de représentation d'informations linguistiques réalisés en différents points du globe puissent être intégrés à peu de frais, et qu'il ne soit pas nécessaire de développer ses propres ressources adaptées au système de distribution qui a été conçu. Les efforts de la communauté internationale en la matière, au travers d'initiatives du type de la TEI (TEI P3, 1994) et dans la même lignée EAGLES, au niveau européen, fournissent un cadre de référence important en la matière ;
- plus largement, il peut s'avérer important de considérer une infrastructure répartie qui permette aux spécialistes de tel ou tel type de ressources de gérer l'origine et la qualité des données qui sont mises à la disposition d'autres chercheurs ;
- la mise en oeuvre doit reposer sur une architecture logicielle ouverte qui puisse intégrer des composants développés par d'autres équipes ou inversement être à même de fournir des modules relativement génériques de gestion de documents par exemple ;
- les utilisateurs potentiels du système télématique doivent être impliqués très tôt dans les phases de conception ainsi que pour toute évolution notable du système. Il est important d'une part que tout utilisateur appréhende correctement la technologie qui lui est fournie (lisibilité de l'interface) et que par ailleurs, les fonctionnalités offertes correspondent à de réels usages existants ou potentiels pour la communauté académique visée (ingénieur, enseignant ou chercheur) ;

- enfin, un tel système doit pouvoir évoluer en fonction des contraintes technologiques nouvelles qui modifient l'environnement de développement et d'accès au service télématique. Nous verrons combien le choix d'une certaine normalisation est aussi important sur ce point.

Plusieurs projets en cours de développement visent à donner accès, via Internet, à des ressources linguistiques. On peut dégager trois grands types d'approches :

- une distribution directe de ressources à l'unité, sans soucis d'accès plus fin au contenu. C'est le cas du LDC (Linguistic Data Consortium) qui fournit des ressources primaires à différents projets de recherche américains ;
- la mise en oeuvre de systèmes spécifiques reposant sur des formats de données et des choix logiciels propriétaires, car reliés — au moins initialement — à un usage très particulier. C'est ainsi que la base Frantext mise en oeuvre à l'INaLF-CNRS a accompagné la réalisation du Trésor de la Langue Française ;
- des systèmes ayant une vocation d'ouverture large vers le monde académique qui, de ce fait, vont s'appuyer au maximum sur des standards existants comme format central de représentation des données. C'est le cas de l'OTA (Oxford Text Archive) dont l'ensemble de la base migre progressivement vers un format conforme aux recommandations de la TEI. C'est aussi dans ce cadre que se place le projet Silfide.

Par ailleurs, de nombreux projets ont vu le jour récemment visant à définir des environnements intégrés de manipulation de ressources électroniques dans la perspective d'intégrer ou au moins de combiner les résultats récents en matière d'ingénierie linguistique (marquage morphosyntaxique, indexation etc.). De façon intéressante, on y retrouve les oppositions relevées pour les systèmes de distribution mentionnés ci-dessus avec des projets qui choisissent résolument un format proche de SGML pour échanger de l'information entre leurs modules (ainsi l'environnement LT NSL - Thompson et McKelvie, 1996 ; ou le langage de requête SGML QL - Harié et alii, 1996 et 1997) et d'autres qui, pour des raisons notamment d'efficacité de traitement, adoptent des formats spécifiques (l'architecture logicielle GATE - Cunningham et alii, 1996 et 1997 ; ou le système ALEP - Simkins, 1994). L'une des difficultés principales à laquelle sont confrontées ces différentes initiatives est bien de combiner une certain respect des normes existantes et de leur logique interne (respect de la structure en arbre dans le cas de SGML par exemple), un réalisme technique suffisant pour aboutir à une implémentation viable (définition simultané d'une API et d'outils spécifiques), et une certaine sensibilité aux

contraintes des applications concrètes de façon à ce que le langage de requête qui a été conçu puisse s'adapter à différents cadres d'usage (possibilité de définir des scripts de requêtes complexes).

On le voit, il s'agit d'un problème complexe qui nécessite à la fois un certain pragmatisme, pour ne pas se laisser noyer par l'ensemble des difficultés à résoudre, et une grande ouverture d'esprit pour suivre au plus prêt les évolutions normatives ou technologiques et ne pas s'enfermer dans des choix qui pourraient s'avérer bloquants.

2. Perspective adoptée au sein du projet Silfide

Silfide (Serveur Interactif pour la Langue Française, son Identité, sa Diffusion, son Etude) est un projet qui, à l'initiative du CNRS et de l'Aupef•Uref, a pour vocation de mettre des ressources linguistiques normalisées à la disposition de la communauté académique francophone. Plus précisément, les missions confiées à Silfide s'articulent autour de trois points principaux :

- mettre en œuvre un véritable *service public* d'accès aux ressources linguistiques, permettant à tout enseignant ou chercheur d'un laboratoire, sans contrainte financière, d'accéder en ligne aux ressources développées au sein d'autres équipes ;
- permettre à des *utilisateurs non spécialistes des outils informatiques* (lexicographe, linguiste, historien etc.) de disposer d'une infrastructure minimale de travail sur des ressources textuelles sans avoir à télécharger des outils qui leur seraient complexes à utiliser ;
- de s'appuyer, autant que faire se peut, sur les normes ou pratiques internationales existantes et promouvoir ces mêmes normes au sein d'une large communauté. C'est ainsi que Silfide a été amené à immédiatement choisir le cadre de la TEI comme support de représentation de l'ensemble des données prises en compte, quelle que soit leur nature (romans, théâtre, articles de journaux, dictionnaires etc.).

Le choix de la TEI comme cadre de référence pour la représentation de nos documents n'est cependant pas un choix aveugle. Il est en effet nécessaire d'effectuer un véritable travail de réflexion éditoriale pour homogénéiser les niveaux de codage des documents accessibles depuis le serveur. En particulier (Romary & alii, à paraître), il nous a fallu définir un canevas précis de champs obligatoires et optionnels au sein de l'en-tête TEI, ainsi qu'un ensemble de recommandations minimales pour les contenus que nous avions à prendre en compte (principalement roman et théâtre dans un premier temps). Ce travail, proche de celui effectué au sein de différents autres projets tels que l'Oxford Text Archive, le Women Writers' Project ou

Parole dans un cadre plus européen, doit bien sûr s'accompagner d'un effort important de concertation à l'image des travaux présentés dans (Burnard & Popham, à paraître).

Du point de vue des utilisateurs, nous avons choisi de développer un scénario qui soit le plus générique possible, en distinguant deux grandes phases. Dans un premier temps, l'utilisateur sélectionne un ensemble de ressources pour se former un corpus de travail. Cette sélection s'opère sur la base d'une recherche multicritère pouvant porter sur différents champs présents dans l'en-tête TEI minimal défini pour nos ressources (titre, auteur, date, genre etc.). Comme l'utilisateur est susceptible d'affiner ses besoins en parcourant la base de documents, le processus envisagé est itératif et repose sur la sélection progressive de ressources au sein d'un « panier » (qu'il peut par ailleurs éditer pour en supprimer des entrées superflues par exemple). Une fois le corpus de travail établi au sein du panier, l'utilisateur active les outils nécessaires à sa recherche, visualisant les résultats ainsi calculés directement sur son navigateur Internet.

Les choix présentés ci-dessus ont conduit à la réalisation d'un premier prototype en ligne (http://www.loria.fr/projets/Silfide) permettant à un utilisateur non répertorié de choisir des ressources au sein d'une base expérimentale d'une centaine de documents et d'y effectuer des concordances élémentaires, contraintes éventuellement par des contextes structurels particuliers. Ce service, par sa disponibilité (gratuité et flexibilité de l'accès via le web), s'est immédiatement imposé à un certain nombre d'équipes de recherche comme cadre de travail pour l'observation de phénomènes linguistique et nous a véritablement permis d'appréhender plus finement les difficultés liés à la mise en place d'un tel système d'accès télématique.

3. Mise en oeuvre d'un réseau de serveurs

L'expérience du premier prototype de serveur Silfide et plus particulièrement les limitations qu'entraînait son architecture logicielle (plate-forme hétérogène combinant des applets Java, des scripts CGI, ainsi que des parties de code écrites en Javascript…) nous ont conduit à mener une réflexion encore plus poussée concernant l'uniformisation des représentations et des traitements, en vue de définir à terme un environnement flexible et générique de manipulation de ressources linguistiques. D'un point de vue théorique, cette réflexion nécessite une remise en compte du document électronique comme une entité monolithique pour étendre cette notion à tout fragment de document, et d'une façon plus générale encore à tout assemblage, physique ou virtuel, de fragments de document. Bien que dans cet article nous ne présenterons que les conséquences immédiates de nos travaux de recherche en la matière sur la définition d'une architecture en

réseau, il est important de garder à l'esprit que l'accès généralisé à des ressources linguistiques doit s'accompagner d'une étude en profondeur sur la nature et les fonctionnalités des objets que l'on souhaite manipuler.

De fait, il est nécessaire d'introduire ici en quelques mots l'un des résultats les plus importants issus des travaux du consortium WWW, le langage de balisage XML, sur lequel reposent plusieurs des idées que nous développons au sein de la nouvelle plate-forme Silfide. Fruit d'une volonté initiale, d'une part, de pallier les insuffisances d'un langage de présentation de l'information tel que HTML (divergence des formats de représentation entre constructeurs, incapacité à structurer les contenus informationnels) et, d'autre part, de proposer un langage qui par sa simplicité de mise en oeuvre puisse être réellement utilisé de façon universelle pour l'échange de documents sur Internet, XML représente maintenant une avancée majeure dans le monde du document électronique. Conçu initialement comme une version simplifiée de SGML, le langage XML intègre un ensemble de fonctionnalités propres qui bouleverse profondément la vision classique du document SGML. On peut ainsi relever les éléments suivants :

- XML repose sur une syntaxe simplifiée qui supprime en particulier bon nombre d'options supportées par SGML (par exemple les indications de minimisation), mais introduit la notion de document bien formé, c'est-à-dire dont on ne vérifie pas la conformité vis à vis d'une DTD[1]. Cette nouveauté permet en particulier de pouvoir manipuler des fractions de document sans avoir à en définir la structure ;
- le langage XML est accompagné d'une proposition précise de deux mécanismes de pointage (Xpointers) et de lien (Xlinks) inspirés des recommandations de la TEI et permettant d'intégrer des références hypertextuelles étendues à l'intérieur d'un document. Cet aspect est essentiel pour tout ce qui touche aux ressources linguistiques car il fournit un cadre de référence pour représenter diverses opérations telles que les annotations externes à un document (e.g. étiquetage morphosyntaxique) ou l'alignement multilingue.

Ces caractéristiques, et notamment la forte compatibilité avec SGML[2], ne font que renforcer le bien-fondé des choix de normalisation adoptés par différents projets, dont Silfide en particulier.

---

[1] Document Type Definition : déclaration d'une structure abstraite de document spécifiant la syntaxe d'utilisation des balises admissibles, ainsi que des attributs associés.

[2] Le format simplifié imposé par la syntaxe de XML est proche de la forme dite canonique que peut produire un parseur classique et de ce fait rend relativement aisé le processus de conversion de documents SGML existants en documents XML bien formés. Le problème est quelque peu plus complexe s'il s'agit d'éditer directement des

Bien plus, la grande flexibilité qu'offre XML permet d'envisager des architectures de systèmes d'accès aux ressources linguistiques encore plus ouvertes et reposant sur des choix de représentation des données qui transitent au sein de ou entre ces systèmes encore plus unifiés.

Le choix d'uniformiser les représentations au sein d'un serveur de ressources linguistiques est le fruit d'un ensemble de réflexions relatives à la logique interne de tels services. Tout d'abord, les données primaires susceptibles d'y être manipulées, même si toutes codées en SGML/XML, sont relativement hétérogènes. Il peut s'agir de textes narratifs, de poésie, de théâtre, ou même de documents plus structurés tels que des dictionnaires mono ou multilingues, correspondant à des DTD de fait différentes[3]. Pourtant, un utilisateur peut très bien voir ces différents documents de façon homogène, pour par exemple interroger uniformément l'ensemble des contenus (e.g. vouloir tous les contextes où le verbe *déplacer* est utilisé de façon transitive, que ce soit dans un paragraphe de roman, un vers de poésie ou un exemple de dictionnaire). En poussant un peu plus loin l'analyse, on peut observer que les résultats d'une requête peuvent eux-mêmes être considérés comme un ou des documents tout aussi bien que les documents primaires dont ils sont éventuellement des extraits. Il est donc important que ces résultats puissent apparaître sous le même format, en l'occurrence des documents XML, ce qui permet de les intégrer dans le « panier » Silfide et donc de les mettre en position d'être eux-mêmes la source d'opérations ultérieures[4].

Nous avons choisi, dans le cadre de Silfide, d'aller encore plus loin dans le sens d'une uniformisation des formats de représentation des informations transitant au sein du réseau de serveurs. Les requêtes sont ainsi représentées sous la forme d'un langage exprimé à l'aide d'un document XML au format XQL (eXtended Query Language) qui intègre d'une part une syntaxe à

---

documents XML sur la base de DTD existantes, qui, suivant leur complexité, sont plus ou moins adaptables au cadre simplifié offert par XML. Ainsi, la DTD simplifiée de la TEI (ou TEI Lite) a déjà été adoptée par plusieurs équipes, alors que la DTD TEI P3 (celle correspondant aux directives complètes de la TEI) utilise un certain nombre de particularités de SGML qui freine son adaptation intégrale au cadre XML. De fait, il n'est pas sûr que réaliser cette opération de transformation soit la bonne réponse à la question si on considère la plus grande flexibilité offerte par XML.

[3] Dans le cadre de la TEI, ces différents types de documents sont contrôlés par une seule DTD où certaines options ont été sélectionnées. Ce qui revient à envisager plusieurs types de documents.

[4] Par exemple, on peut souhaiter filtrer des concordances en y appliquant des contraintes plus strictes, ou effectuer des calculs statistiques de co-occurrence en confrontant des concordances avec le corpus de référence (i.e. le panier initial) dont elles sont extraites.

la SQL et d'autre part un mécanisme d'accès aux documents XML reposant sur une adaptation des pointeurs étendus XML. Un exemple de requête XQL est présenté figure 1. Cette représentation nous permet de parfaitement intégrer la notion de requête dans l'architecture Silfide puisque chaque ensemble résultat peut intégrer la requête dont il est la conséquence, mais aussi de considérer toute requête comme l'expression potentielle de son résultat et de l'intégrer dès sa définition dans l'ensemble des documents manipulés par l'utilisateur. Nous avons ainsi une notion de document virtuel qui permet de gérer d'une part des résultats que l'on ne voudrait pas copier intégralement au niveau de l'utilisateur quand ils forment des quantités trop importantes, ou gérer l'indisponibilité temporaire d'un serveur sur le réseau.

```
<xql>
      <query>
            <select selscrit="all">
            <output>
                  <xptr alias="x1">descendant(all,p)</xptr>
            </output>
            <from>
                  <db url="sample.xml"><xptr id="x1">root()</xptr></db>
            </from>
            <where>
                  <term neg="asis">
                  <grep nag="asis" case="yes">
                        <xptr alias="x1">child(all,s).child(all,#text)</xptr>
                        <regexp>encore</regexp>
                  </grep>
                  </term>
            </where>
            </select>
      </query>
</xql>
```

Figure 1 : une requête simple exprimée à l'aide du langage de requête XQL.

Enfin, toutes les données liées à la gestion de l'espace de travail de l'utilisateur, telles que ses préférences (langue de travail etc.), les caractéristiques de sa session (serveurs sélectionnés) et surtout les données qu'il a sélectionnées dans son ou ses paniers, ainsi que l'ensemble des requêtes et résultats correspondants sont intégrées au sein d'un même document XML qui peut ainsi être facilement manipulé et sauvegardé. Ce document est contrôlé par la DTD SIL (Silfide Interface Language) qui repose sur différentes DTD partielles correspondant à ses principales composantes.

Sur la base des formats de représentation décrits ci-dessus, l'architecture en réseau de la nouvelle plate-forme Silfide prend la forme du schéma de la figure 2. Le client local se connecte à un module d'accueil qui gère les droits d'accès ainsi que l'enregistrement des paramètres de session. Puis, par le biais d'un module de distribution des requêtes qui se connecte sur les

différents serveurs sélectionnés par l'utilisateur (y compris le serveur local), il est possible d'accéder de façon transparente aux bases de ressources distribuées. L'ensemble du réseau fonctionne alors sur la base des principes suivants :

- les utilisateurs sont répertoriés et gérés localement sur un serveur au choix ;
- les requêtes sont transmises de façon transparente à l'ensemble des serveurs sélectionnés, accompagnées du niveau de droits accordé à l'utilisateur ;
- les résultats sont regroupés au niveau de l'espace de travail pour être présentés globalement à l'utilisateur ;
- la gestion du réseau est presque entièrement délocalisée, et seul un module indépendant (non représenté sur la figure) permet d'identifier et de contrôler l'ensemble des serveurs connectés au réseau, ainsi que leurs caractéristiques principales (adresse d'accès, profil des contenus etc.).

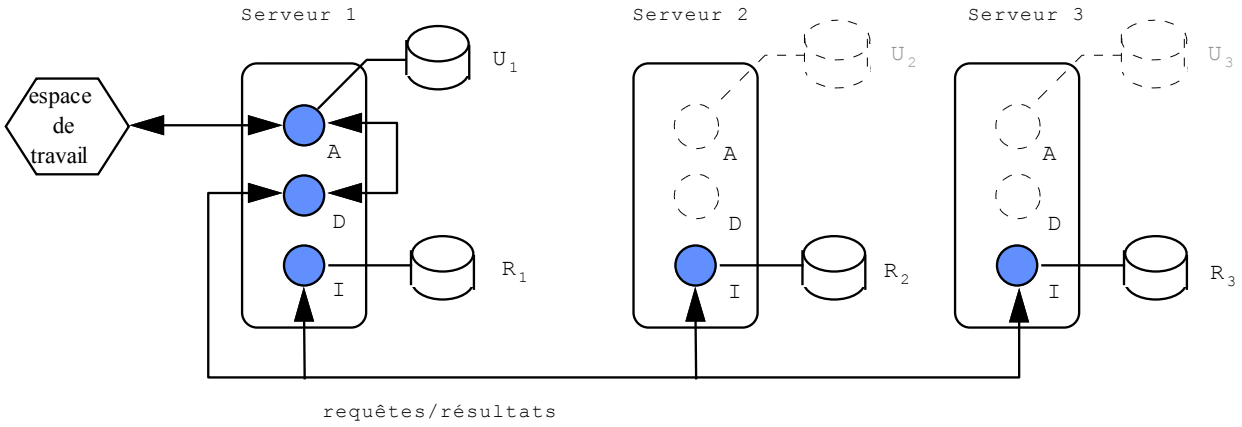

Figure 2 : Architecture générale d'un réseau de serveurs Silfide.

A : accueil ; D : distributeur de requêtes ; I : interpréteur de requête

U : base utilisateur locale ; R : base de ressources linguistiques

L'ensemble des modules associés à chaque serveur a été développé sous la forme de servlets Java indépendantes, ce qui permet une plus grande flexibilité de configuration de chaque serveur.

4. Bilan et perspectives

Le travail mené au cours de la définition du serveur Silfide dans sa première version, puis dans sa configuration en réseau, nous a conduit à réfléchir en profondeur sur la notion même de document structuré et des moyens d'accès qu'il faut y associer. Nous avons ainsi été amené à unifier les représentations associées au différentes informations qui pouvaient transiter au sein d'un réseau

de serveurs de ressources linguistiques de façon à en intégrer les structures logiques (dépendances, relations etc.) et donc les traitements. Il nous semble que nous avons abouti à une structure relativement générique qui devrait pouvoir s'appliquer à des situations similaires d'accès à des bases d'informations semi-structurées, comme par exemple les informations portant sur le génome pour lesquelles émergent en particulier des schémas de structuration s'appuyant sur XML (cf. la proposition BSML, http://visualgenomics.com/sbir/rfc.htm).

Bien que la mise en oeuvre d'un accès télématique à des ressources linguistiques puisse être vue comme une activité de recherche essentiellement appliquée, et effectivement, ces travaux représentent de nombreux mois de travail d'une équipe de plusieurs personnes, il est pour nous l'occasion de faire émerger des pistes de recherche particulièrement intéressantes. Parmi les thèmes en suspens qui nécessiteraient une réflexion approfondie, il serait d'une part nécessaire d'aller plus loin encore dans la définition d'un langage de requête générique pour les documents semi-structurés qui permette de parcourir à la fois des documents primaires et des annotations portant sur ceux-ci (e.g. étiquetage morphosyntaxique, codage de la référence, notes critiques etc.) et d'autre part définir des mécanismes de représentation et de présentation des résultats qui permettent d'intégrer ceux-ci dans une chaîne de consultation et d'édition des documents, là encore dans une perspective d'annotation multiple de documents.

5. Références

CUNNINGHAM (H.), K. HUMPHREYS, (Y.) WILKS et (R.) GAIZAUSKAS 1997, « Software Infrastructure for Natural Language Processing », *Actes Fifth Conference on Applied Natural Language Processing (ANLP-97)*, .

HARIE (S.), (E.) MURISASCO, (J.) LE MAITRE et (J.) VERONIS 1996, « SgmlQL : un langage de requêtes pour la manipulation de documents SGML », *Cahiers de GUTemberg, 24*, pp. 181-184.

HARIE (S.), (J.) LE MAITRE, (E.) MURISASCO et (J.) VERONIS 1997, *SgmlQL : language reference*. (http://www.lpl.univ-aix.fr/projects/SgmlQL/MQL2.html).

BURNARD (L.) et (M.) POPHAM, « Putting our Headers together, a report on the TEI Header Meeting 12 September 1997 », à paraître, *Computers and Humanities.*

ROMARY (L.), (P.) BONHOMME, (F.) BRUNESEAUX et (J.-M.) PIERREL, « Silfide: A System for Open Access and Distributed Delivery of TEI Encoded Documents », à paraître, *Computers and Humanities.*

TEI P3, 1994, Association for Computers and the Humanities (ACH), Association for Computational Linguistics (ACL), and Association for Literary and Linguistic Computing (ALLC) (1994), *Guidelines for Electronic Text Encoding and Interchange* (TEI P3), Editions C. M. Sperberg-McQueen and L. Burnard, 2 volumes, Chicago, Oxford: Text Encoding Initiative.